Contraction Graphs for Representation and Analysis of RNA Secondary Structure

Chris H.Q. Ding[1], Richard F. Meraz[2], Xiaofeng He[1] and Stephen R. Holbrook[2]
[1]Computational Research and [2]Physical Biosciences Divison
Lawrence Berkeley National Laboratory, Berkeley, CA USA
chqding,rfmeraz,xhe,srholbrook@lbl.gov

Abstract

We introduce the contraction graph as a novel graphical representation of RNA secondary structure. A contraction operation – defined along base pars, single stranded regions, and tertiary interactions – allows for representing RNA secondary structure at different levels of detail. Directionality of the graph and flow conservation of the contraction operation allow for explicit representation of the 5' and 3' ends of the molecule. Furthermore, existing representations such as tree and dual graphs are specific instances in this framework. We anticipate that this unified representation will facilitate automated motif finding and functional classification based on direct searches and comparisons of RNA secondary structure topology.

1. Introduction

Secondary structure is critical in defining the shape and function of RNA molecules. Computational methods use the thermodynamics of base pairing and experimentally derived energy parameters [5] to predict secondary structure with reasonable accuracy [6]. Recently, methods to automatically characterize the growing databases [1] of observed and predicted RNA structure have used mathematical representations such as tree and dual graphs [2]. Tree graphs represent RNA loops as vertices and helices as edges to describe connectivity between these featues. Dual graphs represent loops as edges and helices as vertices and are able to represent more topologically complex features such as pseudoknots. Computable representations allow for automated searching, comparison, and enumeration of structures as well as the development of methods for discovering new motifs.

Existing representations alow only one level of secondary structure detail, ignore the natural 5' to 3' directionality of the molecule, and do not incoporate possibly important labelings such as the length of structural features or the identity of individual nucleic acids [2, 4]. We propose a representation of RNA secondary structure in which graphs of different detail or emphasizing different features are generated by a multi-level contraction process on a directed graph.

2. RNA Contraction Graphs

Initially, a given RNA molecule is represented as a directed graph in the 5' to 3' direction. Nodes represent nucleotides and arcs represent covalent links along the phosphoribose backbone. Using a computationally or experimentally derived secondary structure, we contract base paired nucleotides into a single node, conserving arcs in the helical region (Figure 1A). In the second stage, we contract edges of consecutive nodes within helical and single stranded regions such as internal, junction or hairpin loops. In edge contraction, two edges and one node are contracted into an edge which conserves the weights on the arcs (Figure 1B). In the third stage we can obtain tree graphs by contracting adjacent loop residues into a single node, or dual graphs by contracting helical stems into single nodes (Figure 2). Psuedoknots are natural in this representation. During any level of contraction the inbound arc weight must be the same as the out bound arc weight: $N_{in} = N_{out}$. Thus the 5' and 3' ends are easily recognized as the nodes where flow conservation is broken: i.e. $5'\ \mathrm{end} : N_{in} = N_{out} - 1$ and $3'\ end : N_{in} = N_{out} + 1$

3. Advantages and Applications

Our graph representation has a number of advantages over other representations such as tree graphs, dual graphs, and abstract trees [4]. The most important advantage is that, depending on the detail required for a particular problem, there are several levels of abstraction that can be used to represent secondary structure, with the tree and dual graphs as particular instances. In addition, the representation is mathematically consistent and based on a few basic contraction

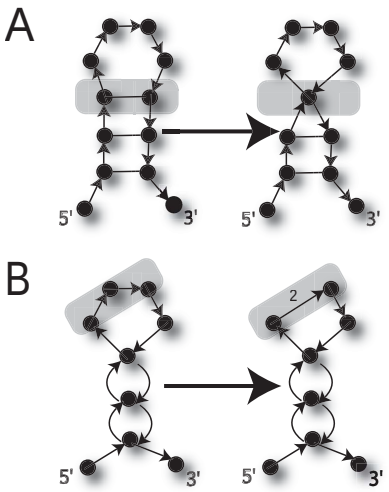

Figure 1. A. The shaded base pair (left) is shown after contraction (right). Arcs are conserved during contraction along base pairs. B. The shaded edges in the loop region (left) are shown after contraction (right). Flow conservation yields a weight of two on the resulting edge. Also note that all base pairs in the helical region have been contracted.

operations which are easily programmed on a computer. Implicit in the representation is directionality based on the 5' and 3' ends of the molecule – an important aspect not used in tree or dual graphs. Beside contraction on covalent or base pairing linkage, contraction can occur using other criteria such as base-stacking, coaxial helices, and other tertiary interactions. Finally there are a number of useful labeling schemes that are compatible with the contraction operation. For example, nucleotide labels could be retained and concatenated for computation of sequence statistics on edges or vertices of the graph. Such a labeling would be useful for global comparisons of topology that include a component based on sequence similarity. Recently we have defined kernels directly on RNA dual graphs to train discriminative classifiers on the families from the RNA Family Database [3] (Karklin Y., Meraz, R.F., Holbrook S.R., unpublished results). This representation is adequate for learning the secondary structure topologies for a variety of families, but suboptimal for others. We anticipate that kernels defined on contraction graphs of varying levels of detail, and with different node and edge labeling schemes that this framework accommodates, will improve the ability of classifiers to learn topological features of RNA families. Other applications include defining contractions that emphasize different secondary structural features to improve search algorithms for various motifs and tertiary contacts.

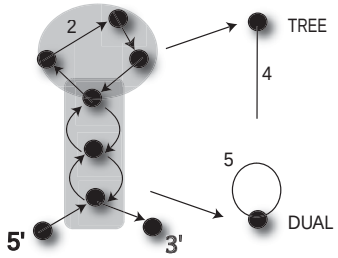

Figure 2. (Left) A partially contracted external loop and a base pair contracted helix are shaded. (Right) Contraction of the loop to a node and the helix to an edge yields a tree graph; while contraction of the loop to an edge and the helix to a node yields a dual graph.

4. Conclusion

RNA contraction graphs can represent different abstractions of secondary structure, explicitly represent the directionality of the molecule, and are compatible with a variety of biologically relevant labeling schemes. The framework includes the existing tree and dual graph representations of RNA secondary structure via contraction operations on a fully expanded graph. Several applications to secondary structure comparison and motif finding are promising.

References

[1] H. H. Gan, D. Fera, J. Zorn, N. Shiffeldrim, M. Tang, U. Laserson, N. Kim, and T. Schlick. RAG: RNA-As-Graphs database–concepts, analysis, and features. *Bioinformatics*, 20(8):1285–91, 2004.

[2] H. H. Gan, S. Pasquali, and T. Schlick. Exploring the repertoire of RNA secondary motifs using graph theory; implications for rna design. *Nucleic Acids Res*, 31(11):2926–43, 2003.

[3] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucleic Acids Res*, 31(1):439–41, 2003.

[4] S. Y. Le, R. Nussinov, and J. V. Maizel. Tree graphs of RNA secondary structures and their comparisons. *Comput Biomed Res*, 22(5):461–73, 1989.

[5] D. H. Mathews, J. Sabina, M. Zuker, and D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5):911–40, 1999.

[6] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–65, 1999.